Praktikum z ekonometrie

VŠE Praha

Tomáš Formánek

Block 6 — Treatment effects analysis

1. Treatment effects: Introduction
2. Differences in differences (DiD) estimator
3. Propensity score matching
4. Regression discontinuity design
5. Treatment effects: self-study materials

Treatment effects: Introduction

- **Treatment effect analysis:** to evaluate the impact of intervention (treatment) on some **outcome** of interest.
- Response to treatment is evaluated relative to a benchmark: no treatment (control) or different treatment.
- Analysis of treatment effect typically involves regression models, with outcome as the dependent variable.
- Treatment effect analysis is generally based on the framework of Rubin's causal models.

Treatment effects: Introduction

Examples

- Wage effect of enrollment in a skill-training program.
- Health effects (speed of recovery), if new drug is used.
- Student performance upon being educated in small classes as opposed to large classes (being in a small class is the treatment).

Key topics of the analysis:
Treatment participation: random assignment or self selection?
Treatment effects: actual effects or influence by confounding factors?

Multiple treatments:
If treatment varies in intensity or type. Same principle of analysis, the choice of benchmark is more flexible.

Treatment effects: Basic notation & terminology

- Every ith individual in a population has a potential outcome y_i and can be exposed to treatment $C_i = \{0; 1\}$.
- $y_{i1} = y_i|(C_i = 1)$ for the treated, and
  $y_{i0} = y_i|(C_i = 0)$ for the non-treated.
- Average treatment effect (averaged across population):

  $$\text{ATE} = E\left[y_{i1} - y_{i0}\right],$$

  but the ith observation only exist in one of the two states.
- **Average treatment effect of the treated** is more of interest:

  $$\text{ATET} = E\left[y_{i1} - y_{i0} | C_i = 1\right],$$

  and the second term y_{i0} is a missing counterfactual.
- Individuals will only exist in one state: treated/untreated. Multiple assumptions apply for ATE/ATET estimation.

Treatment effects: Study & data types

Types of studies and data used for TE analysis:

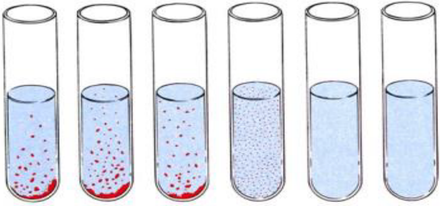
- Scientific experiments under randomisation: assignment into treated and control groups is random. Relatively rare in socio-economic studies.
- Observational studies (natural experiments, quasi-experiments): assignment into treatment and control group is not random.

Main problems of TE analysis:

- **Endogeneity of treatment** and self-selection bias: if treatment participation is optional, individuals who choose to participate may be systematically different from non-participants.
- **Missing counterfactual:** Individuals always measured as either treated or untreated, we cannot observe both y_{i1} and y_{i0}.

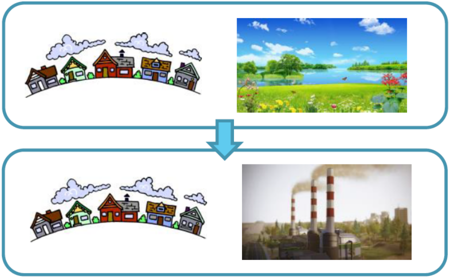
Treatment effects: Study & data types

Scientific experiment



- Test tubes identical except for catalyst
- Measure: Effect at different catalyst volumes (reaction speed, product volume, . . .)
- Perform the experiment n-times
- Control for other factors (heat, . . .)
- Estimate average effects & standard errors

Natural experiment (quasi-experiment)



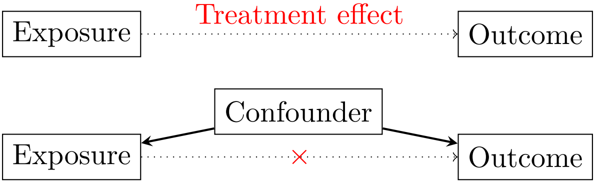
- Garbage incinerator is built in one given suburban area over time
- How do we estimate the effect on individual house-prices?
- Identical control group does not exist. . .
- Different estimators exist – multiple assumptions apply!

Treatment effects analysis & natural experiments

Three main types of analysis, key assumptions & requirements:

- **Differences in differences (DiD) estimator**
  - Independence of treatment on the outcome at the base (y_{i0}), i.e. treatment assignment exogenous/random
  - Parallel trends assumed
- **Propensity score matching (PSM):**
  - Unconfoundedness (treatment assignment as good as random after accounting for covariates)
  - Common support (for the treated and untreated)
  - Large sample
  - PS is essentially the probability of treatment participation
- **Regression discontinuity design (RDD):**
  - Assignment variable & assignment threshold exist
  - Local randomization assumed

Treatment effects: Unconfoundedness



Often: combination of actual TE & confounding factor influence.

Example: Company has two types of trucks, A and B (using the new A-Truck is the treatment). We want to compare fuel efficiency (outcome). Using MPG, we find A-Trucks are more efficient. However, A-Trucks often drive on highways, while B-Trucks are mostly in the city. This route difference is a confounding variable, making our results unreliable as highway driving is generally more fuel-efficient.

To improve the study, we could randomize truck assignments for equal city and highway driving, introduce city driving as another independent variable to a model, or segment the study into city and highway driving comparisons.

Treatment effects: Unconfoundedness

Unconfoundedness states that the potential outcomes are independent of the treatment assignment, conditional on a set of observed covariates. Once we control for these covariates, the treatment *assignment* does not provide any additional information about the *potential* outcomes:

$$(y_{i0}, y_{i1}) \perp C_i \,|\, \boldsymbol{X}_i \,, \quad \text{for } \forall i,$$

- (y_{i0}) and (y_{i1}) are potential outcomes under control and treatment,
- C is the treatment assignment (dummy variable),
- $\boldsymbol{X}$ is the set of observed covariates.

Independence of treatment on the outcome at the base (for the untreated) is a stronger assumption. Here, the potential outcome under control is independent of the treatment assignment, even without conditioning on covariates:

$$y_{i0} \perp C_i \,, \quad \text{for } \forall i.$$

Treatment effects: Unconfoundedness

Unconfoundedness assumption allows for the possibility that treatment assignment might be related to the potential outcomes through the observed covariates. Once we control for these covariates, the treatment assignment is assumed to be as good as random.

Independence of treatment on the outcome assumption requires the treatment assignment to be as good as random without conditioning on any covariates. This is a stronger assumption and is less likely to hold in real-world situations.

Treatment effects: Unconfoundedness & do() notation

$$P(Y|\texttt{do}(X)) = P(Y)$$

LHS can be red as: “the probability of Y, given that you do X”. The expression above describes the case where Y is independent of doing X.

Say, we have an outcome y, a dummy treatment variable C and potentially counfounding variable z that influences both y and C. If unconfoundedness holds, the following equation holds

$$P(y|\texttt{do}(C)) = P(y|C)$$

for all values C_i and y_i, where $P(y_i|C_i)$ is the conditional probability. This equality states that C and y are not confounded whenever the observationally witnessed association between them is the same as the association that would be measured in a controlled experiment, with C randomized. Otherwise, we have to account for the confounding factor z:

$$P(y|\texttt{do}(C)) = \sum_{z} P(y|C,z)P(z).$$

https://en.wikipedia.org/wiki/Confounding

Treatment effects analysis

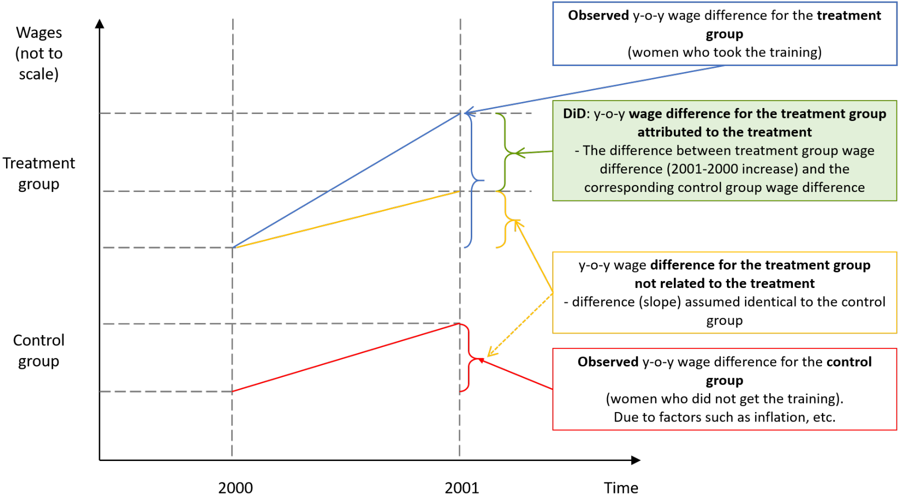
Estimation approaches for TE analysis:

1. Differences in differences (DiD)
2. Propensity score matching (PSM)
3. Regression discontinuity design (RDD)

DiD: Introduction & example

Treatment: employee training for women returning from maternal leave

Outcome: wage effect



Observed y-o-y wage difference for the **treatment group** (women who took the training)

DiD: y-o-y **wage difference for the treatment group attributed to the treatment**
- The difference between treatment group wage difference (2001-2000 increase) and the corresponding control group wage difference

y-o-y wage **difference for the treatment group not related to the treatment**
- difference (slope) assumed identical to the control group

Observed y-o-y wage difference for the **control group** (women who did not get the training). Due to factors such as inflation, etc.

DiD estimator: Assumptions

- *Exchangeability (parallel trends)*: in the absence of treatment, the average outcomes for the treated and control groups would follow the same trend over time.
- *Positivity*: every individual (unit) has a positive probability of receiving the treatment.
- *Stable unit treatment value assumption (SUTVA)*: the potential outcomes for any individual do not depend on the treatment assignment of the individual, nor of the other individuals.
- *No spillover effects*: the treatment of one individual does not affect the outcome of another individual.
- *Treatment unrelated to outcome at baseline*: the allocation of treatment is not determined by the outcome.
- *Stable composition of treatment and control groups*: applies to repeated treatments/interventions.

DiD: Model motivation

With cross-sectional data and **exogenous** treatment variable $C = \{0; 1\}$, we can formulate (estimate, interpret) a regression model such as:

$$y_i = \boldsymbol{x}_i' \boldsymbol{\beta} + \delta C_i + \varepsilon_i.$$

DiD estimator refers to panel-based (or using pooled CS data) analysis, based on two dummy variables and their interaction: T differentiates between two time periods (before/after treatment) and C distinguishes the two groups (treatment/control):

$$y_{it} = \beta_0 + \beta_1 C_i + \beta_2 T_t + \delta_1 (C_i T_t) + \boldsymbol{x}_i' \boldsymbol{\beta} + \varepsilon_{it}.$$

DiD estimator: Model description

$$y_{it} = \beta_0 + \beta_1 C_i + \beta_2 T_t + \delta_1 (C_i T_t) + \varepsilon_{it},$$

where:

- $i = 1, \ldots, N; \quad t = 1, 2.$

T_t is a time dummy, $T_1 = 0$ is the first (pre-treatment) period and $T_2 = 1$ is the second period (post treatment),

C_i is a treatment dummy, $C_i = 1$ for the treated,

$C_i T_t$ is an interaction element, i.e. $(C_i \cdot T_t)$,

δ_1 is the DiD estimator (coefficient).

For simplicity, the term $\boldsymbol{x}_{it}'\boldsymbol{\beta}$ is removed here. Note that adding $\boldsymbol{x}_{it}'\boldsymbol{\beta}$ back to the model doesn't change the general interpretation of δ_1.

DiD estimator: Interpretation of the DiD δ_1 coefficient

$$y_{it} = \beta_0 + \beta_1 C_i + \beta_2 T_t + \delta_1 (C_i T_t) + \varepsilon_{it}$$

Table: Illustration of the DiD estimator

$E(y_{it} \mid C_i, T_t)$	Before $(t=1)$	After $(t=2)$	After – Before
Control $(C_i = 0)$	β_0	$\beta_0 + \delta_0$	δ_0
Treatment $(C_i = 1)$	$\beta_0 + \beta_1$	$\beta_0 + \delta_0 + \beta_1 + \delta_1$	$\delta_0 + \delta_1$
Treatment – Control	β_1	$\beta_1 + \delta_1$	δ_1

Again, if $\boldsymbol{x}_{it}\boldsymbol{\beta}$ is added back to the equation, interpretation of δ_1 remains essentially unchanged.

DiD estimator: Interpretation of the DiD δ_1 coefficient

$$y_{it} = \beta_0 + \beta_1 C_i + \beta_2 T_t + \delta_1 (C_i T_t) + \varepsilon_{it},$$

By rearranging an estimated model, we can express δ_1 as follows:

$$\hat{\delta}_1 = (\overline{y}_{Tr,\,t=2} - \overline{y}_{Co,\,t=2}) - (\overline{y}_{Tr,\,t=1} - \overline{y}_{Co,\,t=1}),$$

which may be also rearranged as:

$$= (\overline{y}_{Tr,\,t=2} - \overline{y}_{Tr,\,t=1}) - (\overline{y}_{Co,\,t=2} - \overline{y}_{Co,\,t=1}),$$

where
Tr subscript stands for treatment group, and
Co subscript identifies the control group.

DiD estimator: Example

What is the effect of building garbage incinerator on housing prices?

```
Dependent Variable: RPRICE
Included observations: 321

       Variable          Coefficient    Std. Error    t-Statistic     Prob.

      (Intercept)          82517.23      2726.910      30.26034       0.0000
          Y81              18790.29      4050.065      4.639502       0.0000
        NEARINC           -18824.37      4875.322     -3.861154       0.0001
     Y81*NEARINC          -11863.90      7456.646     -1.591051       0.1126

R-squared                  0.173948     Mean dependent var           83721.36
Adjusted R-squared         0.166131     S.D. dependent var           33118.79
S.E. of regression         30242.90     Akaike info criterion        23.48429
Sum squared resid          2.90E+11     Schwarz criterion            23.53129
Log likelihood            -3765.229     Hannan-Quinn criter.         23.50306
F-statistic
Prob(F-statistic)
```

RPRICE - house price in real terms (USD)
$Y81$ – dummy variable for 1981, $(t = 1978, 1981)$,
1978 – before “rumors”; 1981 – incinerator operational
NEARINC – dummy for the treatment group

DiD estimator: selection bias

Selection bias (treatment effect vs. selection bias):

Incinerator example: Say, we have a “poor neighborhood” with relatively old and small houses and low house-prices. For complex reasons, it suffers from a representation deficit within the local city council (as compared to other “rich neighborhoods”) and is therefore more likely to get the incinerator. To address this problem, we would need variables to control for this factor.

Job training example: For a natural experiment with job-training effects, we have voluntary participation. If more agile workers tend to participate more, we cannot assume treatment and control group are identical – except for the treatment. Hence, besides of treatment effect, DiD would reflect latent “propensity” to participate.

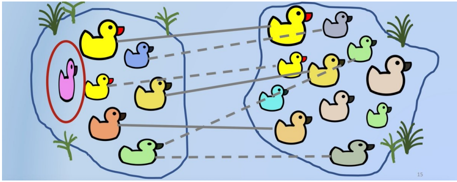
DiD will be biased and inconsistent in both cases.

Treatment effects analysis

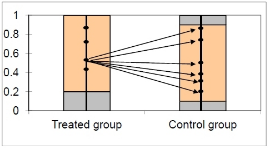
Estimation approaches for TE analysis:

1. Differences in differences (DiD)
2. Propensity score matching (PSM)
3. Regression discontinuity design (RDD)

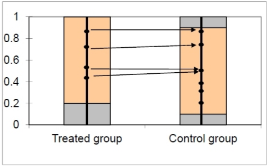
Propensity score matching: Matching vs. PSM



Kernel matching



Nearest neighbor matching



PSM: Assumptions and features

- *Unconfoundedness:* assignment to the treatment is independent of the potential outcomes, given a set of observed covariates.
- *Common Support* or overlap condition, meaning that individuals with the same characteristics (PS) have a nonzero probability of being in the treatment group and in the control group.
- *Large sample sizes:* PSM often requires large sample sizes because matching may not be possible if the overlap between treatment and control groups is not sufficient.

- *Matching*: PSM involves pairing treated and untreated subjects who have similar propensity scores (PS – estimated probability of treatment participation).
- *Propensity score* is a balancing score. Conditional on the propensity score, distribution of observed baseline covariates will be similar between treated and untreated subjects.

PSM: Main steps

1. Collect data, identify that PSM is viable and appropriate
   - Check assumptions
   - Basic analysis using non-matched data
2. Calculate propensity scores
   - Logit/probit with exposure (treatment) as dependent variable
   - Include all predictors of the exposure and none of effects of the exposure. Confounders and interaction variables can be included.
3. Match subjects on the propensity scores
   - 1-to-1 matching, 1-to-n matching
   - Nearest neighbor, Caliper (type of NN), Kernel, Radius, etc.
4. Assess quality of the matching
   - Substantial overlap in covariates between treated & control groups
   - Use diagnostic metrics (e.g. standardized difference)
5. Analyze the propensity-matched data
   - Multiple regression models (use treatment as model variable)
   - DiD on matched data, survival analysis, etc.

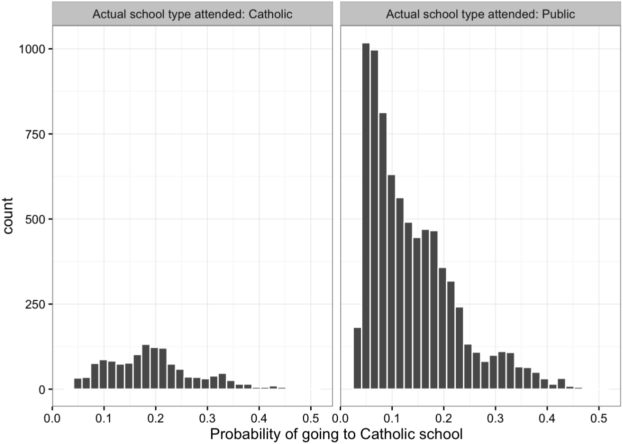
PSM-based analysis

1. Basic analysis on non-matched data – Difference-in-means for:
    - Outcome variable (before & after treatment)
    - Regressors (covariates)
    - Regressors (before & after treatment)

2. Propensity score calculation
    - Estimate the propensity score by logit (probit) model
    - Predicted values $\hat{y}_i$ (on the response scale) are the propensities to treatment (expected probabilities of treatment participation)
    - Evaluate common support (e.g. by plotting histograms of estimated propensities by treatment status)

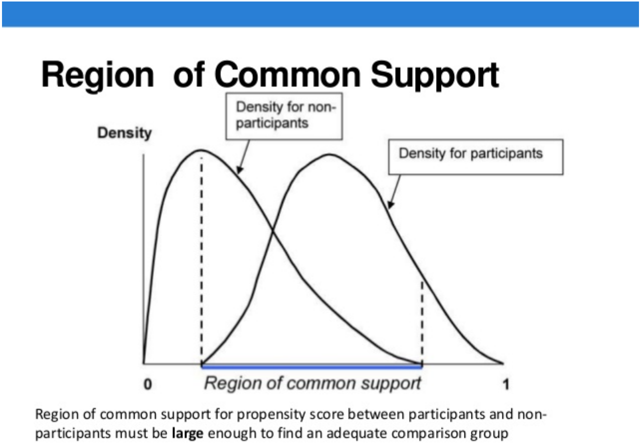
PSM-based analysis

PS – Common support evaluation example
(Going to catholic school is the treatment)



PSM-based analysis

PS – Common support evaluation example
(General case)



Region of Common Support

Region of common support for propensity score between participants and non-participants must be **large** enough to find an adequate comparison group

PSM-based analysis

Simple method for estimating the treatment effect:
(based on propensity scores)

- Restrict the sample to observations within the region of common support (based on estimated PS values).
- Divide the sample within the region of common support into 5 quintiles, based on the estimated propensity scores.
- For each of these 5 quintiles, estimate the mean difference in outcome variable by treatment status.
- Rubin and others suggest that this is sufficient to eliminate 95% of the bias due to confounding of treatment status with a covariate.

PSM-based analysis

3 **Propensity score-based matching**

We seek to find pairs of observations that have very similar propensity scores, but that differ in their treatment status.

Many PSM-based methods exist, we focus on common approaches that fit most types of data.

- **Nearest neighbor:** for each ith treated person, the jth untreated counterpart with the closest propensity score is selected as the match using the expression (with/without replacement):

$$\min_j |\hat{y}_i - \hat{y}_j|,$$

where $\hat{y}_i$ is the PS for the ith participant (treated) and $\hat{y}_j$ is the PS of the jth non-participant (untreated).

PSM-based analysis

- **Caliper** is a variation of nearest neighbor matching: the jth non-participant is only selected as a match to the ith participant, if the PS distance is within the caliper limit

  $$|\hat{y}_i - \hat{y}_j| < \varepsilon,$$

  where the typical value of the tolerance is $\varepsilon = 0.25\sigma_{\hat{y}}$, i.e. it is given as $\pm\frac{1}{4}$ st.err. (of the estimated propensity scores).

  Variants of caliper-based matching:

  - 1-to-1 using NNs within caliper (typical application)
  - 1-to-1 using Mahalanobis within caliper
  - 1-to-n using NNs/Mahalanobis within caliper
  - selection with / without replacement
  - non-treated matches can be based on actually observed individuals or synthetic (combined) control group members, etc…

PSM-based analysis

- **Mahalanobis & caliper:** For each ith participant, we search for the 1 (or n) nearest available matching non-participants, based on Mahalanobis metric, within the i-specific caliper. Mahalanobis distance (metric) is defined as

  $$d(i,j) = (\boldsymbol{u}_i - \boldsymbol{v}_j)'\boldsymbol{\Sigma}^{-1}(\boldsymbol{u}_i - \boldsymbol{v}_j),$$

  where $\boldsymbol{u}_i$ and $\boldsymbol{v}_j$ are vectors of selected matching variables for a given participant i and for some jth non-participant (within caliper limit), and $\boldsymbol{\Sigma}$ is the sample variance-covariance matrix of the matching variables from the full set of non-participants.

  Mahalanobis distances are calculated and evaluated for all $\{i,j\}$ pairs within the caliper for the ith participant.

PSM-based analysis

4 **Assessing quality of the matching**

- Evaluate common support in covariates (follows the same logic as for the common support in PS)
- Visual inspection of covariates (treated vs untreated)
- t-tests of difference-in-means for covariates (treated vs untreated)
- Calculate and assess the average absolute standardized difference (standardized imbalance)

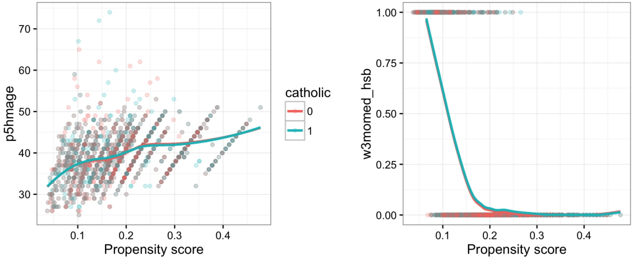
PSM-based analysis

- Visual inspection of covariates (treated vs untreated)

  Plot the mean of each covariate against the estimated propensity score, separately by treatment status.

  If matching is “done well”, the treatment and control groups will have (near) identical means of each covariate at each value of the propensity score (recall the large sample assumption for PSM).

PSM – Common support in covariates example



PSM-based analysis

- Average absolute standardized difference (standardized imbalance)

  A popular measure of of the standardized average imbalance is the following metric:

$$SAI = \frac{1}{K} \sum_{k=1}^{K} \frac{|\beta_k|}{\sigma_k},$$

  where $|\beta_k|$ is the absolute value of difference between the kth covariate means in the treated and control groups in the matched sample, K is the number of covariates and σ_k is the standard deviation of the kth covariate in the matched sample.

  An average absolute standardized difference that is close to 0 is preferable, since that indicates small differences between the control and treatment groups in the matched sample.

PSM-based analysis

5 **Analyzing propensity-matched data**

Estimating TE is simple, once we have a matched sample that we consider well balanced.

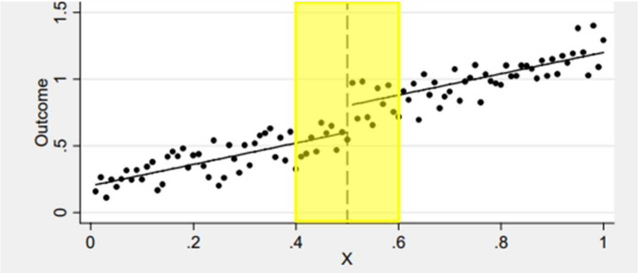
- We use a *t*-test on means of the outcome variable (treated vs untreated)
- We use OLS: SLRM with outcome variable regressed on the treatment variable
- We use OLS, expand model by covariates
- DiD on matched sample (OLS-based analysis)
- Apply any model / analysis tool of choice to the matched data.

Treatment effects analysis

Estimation approaches for TE analysis:

1. Differences in differences (DiD)
2. Propensity score matching (PSM)
3. Regression discontinuity design (RDD)

RDD-based treatment effect visualization example
(linear function & common slopes assumed).



RDD: Assumptions and features

- *Assignment Variable*: RDD requires a measurable assignment variable that determines whether an observation falls into the treatment or control group.
- *Threshold*: specific value of the assignment variable that determines the assignment to treatment (deterministic and fuzzy variants of RDD exist).
- *Local Randomization*: Around threshold, the assignment to treatment is as good as random, which allows for causal interpretation of TE.
- *Continuity*: The expected outcome in the absence of treatment must be a continuous function of the assignment variable at the threshold.
- *Bandwidth Selection*: Bandwidth around threshold needs to be carefully chosen to balance bias and variance.

- RDD estimates are local to the threshold and may not generalize to the whole population – unlike PSM, which gives ATET (PSM-based TEs relate to the treated and assume that TEs are constant across all individuals).
- RDD requires fewer data assumptions and can be used with fewer observations (compared to PSM)

RDD: description of main features

Deterministic / sharp RDD

- There is a continuous variable X_i that determines who gets treated, ($C_i = 1$ if treated).
- X is called the running / assignment variable.
- In sharp RDD, a unit is treated if $X_i \geq c$ and not treated if $X_i < c$, with c being a given threshold.
- We must observe X and know the cutoff/threshold value c.

Fuzzy RDD

- In fuzzy RDD, C_i is a random variable, influenced by X.
- $E[C_i|X_i] = P[C_i = 1|X_i]$ is discontinuous at $X_i = c$.
- X with values close to c (at c) is a predictor of treatment, but it does not completely determine treatment assignment.

RDD: description of main features

With sharp RDD, we can write the following limits
(note the direction of arrows):

- $\lim_{x \to c} E[y_i | X_i = x, C_i = 0] \approx E[y_{i0} | X_i = c]$,
- $\lim_{x \leftarrow c} E[y_i | X_i = x, C_i = 1] \approx E[y_{i1} | X_i = c]$,

where y_{i0} and y_{i1} relate to the potential outcome with/without treatment and the difference between the two expected outcomes at $X_i = c$ is the treatment effect (TE evaluated at $X_i = c$).

In practical applications, RDD-based TE evaluation is calculated for a specific X-value range around c (bandwidth selection).

RDD: description of main features

RDD: Variance vs Bias

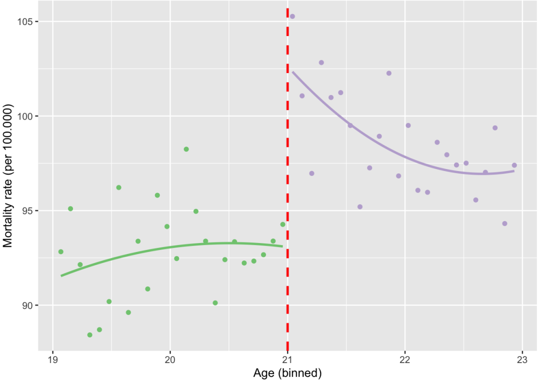
- TE estimated based on small subsample – the value of the running variable is close to c. Smaller sample → larger st. errors.
- Select a larger sample and estimate parametrically (e.g. use regression model). Approach depends on the functional form and polynomials used in model.
- Choice of model specification is complicated by lack of overlap.

Sharp RDD: Lack of overlap

- Overlap requires $0 < P(C_i = 1|X_i) < 1$ for the domain of X_i
- Sharp RDD: $P(C_i = 1|X_i < c) = 0$ and $P(C_i = 1|X_i \geq c) = 1$
- We rely on extrapolation (counterfactuals) to estimate TE.
- TE can be wrong, if we use a wrong functional form $y_i = f(X_i, \boldsymbol{x}_i)$.
- We never know the actual $f(\cdot)$. Model specification is a key issue: use non/semi/parametric method.

RDD: description of main features

RDD-based treatment effect visualization example (quadratic function & separate slopes assumed).



RDD: estimation algorithm

1. Check the data and assumptions (e.g. for sharp RDD)
2. Estimate linear model, start with common slopes
   Calculate TE (make bandwidth selection around c).
3. Include polynomials, allow for separate slopes
4. Modify functional form: use semi/non parametric estimation, splines, LOESS (locally estimated scatterplot smoothing), etc.
5. Sensitivity analysis (compare results based on functional form, bandwidth, etc.)

With fuzzy RDD, TE estimation can follow the IV (instrumental variable) approach. We use the assignment/running variable as an IV. (see lecture notes for RDD here)

Treatment effects – additional literature

Treatment effects analysis

For detailed & technical discussion of TEs, see:

1. Greene: Econometric analysis, chapter 19.6
2. Angrist, Pischke: Mostly Harmless Econometrics
3. Cameron, Trivendi: Microeconometrics, Methods and Applications, chapter 25
4. Wooldridge: Econometric analysis of C-S and panel data, chapter 21 Estimating Average Treatment Effects

- Health Services Research Methods I (HSMP 7607), UC Denver
- https://simonejdemyr.com/r-tutorials/statistics
- https://rpubs.com/sharmaar/RDD